Automating After Action Review:
Attributing Blame or Credit in Team Training

Jonathan Gratch
University of Southern California
Institute for Creative Technology
13274 Fiji Way, Marina del Rey, CA 90292
gratch@ict.usc.edu

Wenji Mao
University of Southern California
Institute for Creative Technology
13274 Fiji Way, Marina del Rey, CA 90292
mao@ict.usc.edu

ABSTRACT: *This paper presents automated methods for facilitating after action review in team training exercises. Much of the learning from team training arises from frank after-the-fact discussions of the exercise, combining individual attributions of blame or credit into a more objective view of what transpired. These individual attributions are social judgments involving not only causality but also explanations of individual responsibility, free will and mitigating circumstances. Such judgments are a key aspect of social intelligence and underlie social planning, social learning, natural language pragmatics and computational models of emotion. Here we introduce a computational model of this judgment process based on psychological Attribution Theory and discuss its potential to facilitate after action review in team training.*

1 Introduction

This paper addresses an important issue in team training: the problem of giving credit where credit is due when multiple parties are involved in a training exercise. Many training applications such as military training or urban disaster simulations involve distributed decision-making and distributed knowledge. No one individual has complete authority or complete access to information. The shared nature of group training considerably complicates the problem of generating feedback to individuals on where they are performing well, where they need improvement, and whether perceived failures are actually outside the individual's control. In this article we intro-duce automated techniques that facilitate credit assignment in team training simulations.

Team training simulations provide individual procedural training during the course of the exercise, but much of the learning occurs after the fact, through the process of *after action review* (AAR). This is an after-the-fact discussion of an exercise where participants can discuss events from multiple perspectives and come to an individual understanding of what happened, why, and how to sustain strengths and improve weaknesses. Bring together contrasting explanations is often the key element in AAR. In a team exercise, individuals view success and failure through the lens of their individual perceptions and decisions, but given the distributed nature of these exercises, an individ-

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE **2003**	2. REPORT TYPE	3. DATES COVERED **00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Automating After Action Review: Attributing Blame or Credit in Team Training	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of California,Institute for Creative Technologies,13274 Fiji Way,Marina del Rey,CA,90292**	8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**10**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Figure 1: Interaction between the trainee and autonomous characters in the Mission Rehearsal Exercise

ual's perspective is frequently biased or misleading. For example, what seems like a bad decision from the perspective of a subordinate might make more sense if he were placed in his commander's shoes. By collectively discussing events after the fact, decision makers can come to a truer understanding of who did what, why, and how they can do better the next time.

Underlying AAR is the problem of forming *social* explanations. In reflecting on an exercise, individuals must form judgments not only of causality but individual responsibility, free will and mitigating circumstances. Did an individual choice cause a significant outcome? Were they simply following orders? Did they agree with the orders and, if not, was this disagreement communicated to their superiors? By eliciting such social explanations, individuals can move beyond their gut feelings about the exercise and contrast and compare the factors underlying these attributions.

Facilitating an effective AAR is an art and it is particularly difficult in large training simulations. Effective AAR demands an expert facilitator that has a sense of the key events in the exercise and can guide and focus the discussion. It also demands that participants can reflect on their individual attributions of blame and credit and relate the underlying factors that led to them. Neither demand is easily met. For example in military exercises such as the Simulated Theater of War (STOW), thousands of entities involved. Decisions are being made at multiple levels in the command hierarchy, and many decision-makers are autonomous or semi-autonomous agents with opaque reasoning processes. In such exercises, facilitators can be easily overwhelmed by the complexity of the exercise and many of the key decision-

makers (e.g. the autonomous agents) are incapable of participating in the AAR process.

This paper lies out a preliminary computational approach to forming social explanations based on psychological attribution theory. We see several immediate applications of this model. Here we focus on its potential to facilitate the after action review process. We see two complementary vectors to apply this technology to AAR. First, these techniques can assist an AAR facilitator by analyzing features of a simulation and communication between individual decision-makers, and identifying key decision events where a different decision or better information may have led to a better outcome. Second, by incorporating social credit attributions into autonomous agents, these agents can form social judgments from their own perspective and potentially participate in the AAR process, for instance, by answering questions on whom they blame and why.

2   Motivating Example

We have been developing these techniques in the context of an Army leadership-training simulator [Rickel *et al*., 2002], and an example from this system illustrates the social factors involved in judgments of blame and credit. The Mission Rehearsal Exercise (MRE) is a virtual reality training environment designed to teach decision-making skills in high-stakes social situations. Intelligent agents control characters (virtual humans) in the virtual environment, playing the roles of locals, friendly and hostile forces, and other mission team members. The goal is to support realistic face-to-face interactions, requiring an emphasis on creating "broad agents" that integrate motor skills, problem solving, emotion, gestures, facial expressions, and language. The virtual humans engage in task-oriented reasoning, and communicate through verbal and non-verbal behavior, including emotional responses. The goal is to support training the social and human-centered aspects of command decision making.

Although the MRE has currently focused on small unit operations, it embodies essential features of many group decision-making problems, including the fact that authority, decision-making responsibility, and perception are distributed across a group of individuals. In the simulation, the trainee is placed in command of an infantry platoon, Eagle 2-6, supporting peacekeeping operations near the Bosnian city of Celic. The trainee's mission is to reinforce another unit, Eagle 1-6. In route, one of his vehicles seriously injures a civilian, and the international press is already on scene. The trainee must balance whether to continue the mission or render aid. Many decisions and outcomes are possible. In the following example, the trainee decides to split his forces, ordering his sergeant to send half of his squads to aid the other unit. His sergeant responds that this is a bad idea; it will allo-
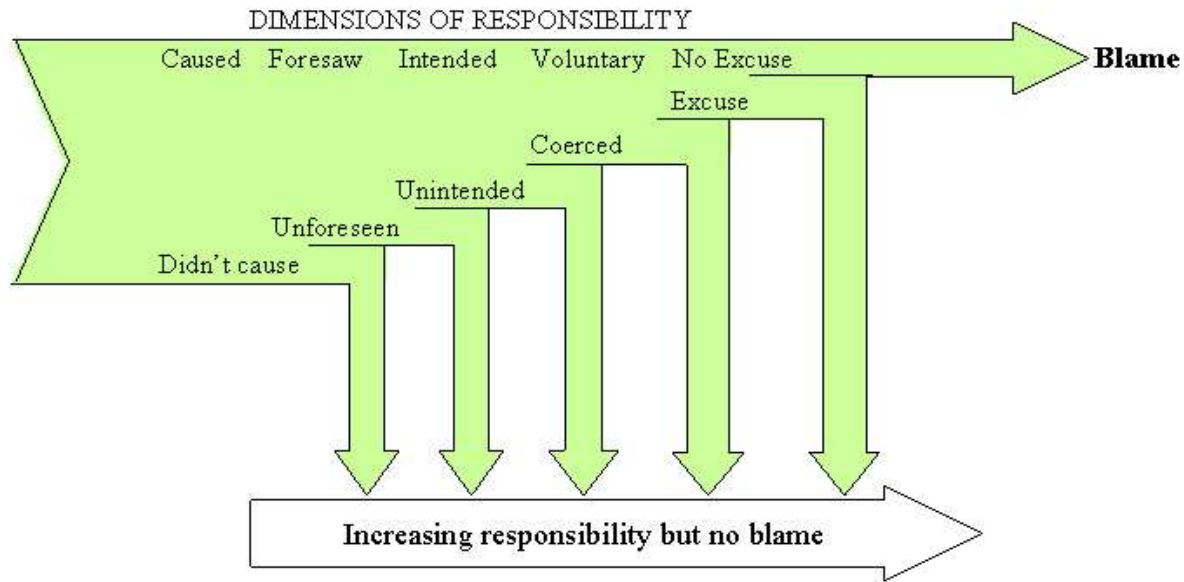
DIMENSIONS OF RESPONSIBILITY

Figure 2: Shaver's Attribution Model of Blame, simplified and adapted from [Shaver 1985].

cate insufficient forces to either goal, and instead, one squad should be sent ahead to scout the route. The trainee overrules this recommendation, restating the original order. In the end, the trainee finds he has insufficient resources to render aid in a timely manner, a fact that is duly noted in the nightly news. The central question addressed here is to assess who, if anyone deserves blame for this unfortunate outcome. Did the trainee truly have a more effective alternative? Did he have sufficient information to make an effective decision? Was the problem with the order or how it was carried out?

Individuals may differ in whom they praise or blame in a specific situation, but psychologists agree on the broad features individuals use to make such judgments. Did the agent *cause* the outcome? Did he/she *intend* the act? Did he/she *know* the consequence? Did he/she have *choice* or was he/she *coerced* by another agent? In the example, we may infer from the conversation that there were alternatives and the sergeant was coerced by the trainee to follow an undesirable course of action. We can further surmise that the trainee must have known the consequence since his sergeant said so. Baring unknown mitigating factors (e.g. the sergeant always gave bad advice in the past), we would likely conclude that the trainee is to blame for the delay and negative press. This example shows that proper assignment of credit or blame in a social setting must not only consider the actions (both physical acts and speech acts) and knowledge state of different actors, but also need to make use of information available to reason about key attributions that contribute to the judgment process.

3 Attribution Theory for Social Judgment

The assignment of social credit or blame has been studied extensively in philosophy, law, and social psychology. As our primary goal is to inform the design of training simulations that model human behavior [Gratch *et al.*, 2002], our models focus on descriptive rather than proscriptive models (i.e., what people do rather than what they should do). In contrast, much of the related work on AI has focused on trying to identify "ideal" principles of responsibility (e.g. the legal code or philosophical principles) and ideal mechanisms to reason about these, typically contradictory principles (e.g., non-monotonic or case-based reasoning) [McCarty, 1997].

In modeling the assignment of social credit, we build on psychological attribution theory, specifically the work of Weiner [1995] and Shaver [1985] as they are readily adapted to artificial intelligence knowledge representation and reasoning methods. Shaver's model is illustrated in Figure 2. In these models, the assignment of credit/blame is a multi-step process that is initiated by events with positive or negative consequences. First one assesses *causality*, distinguishing between personal versus impersonal causality (i.e. whether the consequence is caused by a human or by environmental factors). If caused by a human, judgment proceeds by assessing key factors: was it the actor's *intention* to produce the outcome; did the actor *foresee* its occurrence; was the actor forced under *coercion* (e.g., was the actor acting under orders)? As the last step of the process, proper degree of credit/blame is assigned to the responsible agent. Causal-

ity, intention and foreseeability map to standard concepts in agent-based systems, particularly frameworks that explicitly represent beliefs, desires and intentions [Bratman, 1987; Pollack, 1990; Grosz and Kraus, 1996]. Coercion requires some representation of social relationships and some understanding of the extent to which the coercion limits one's range of options. For example, one may be ordered to carry out a task but to satisfy the order, there may be many alternative ways that vary blame or credit-worthiness.

A system to support the AAR process based on attribution theory must be able to rapidly scan a trace of simulation events to assess instances of causality, intentionality, foreknowledge and coercion. In a standard AAR process, this knowledge comes from asking people to elicit their mental state at the time of certain simulation events. When a simulation involves computer-generated decision makers, it may be possible to encode such information directly in the simulation trace [Johnson, 1994]. In general, however, we would like to model the process that people generally use in attributing credit/blame, rather than assume the perceiving agent has privileged access to other's mental state. In human social interactions, such attributions are gleaned from a variety of sources: from observation of behavior, from statements made through natural language, from knowledge and models built up through past interactions, stereotypes, and cultural norms. We show how to infer such information by analyzing communication traces between entities and by making use of agents' knowledge of actions and consequences and commonsense intuition.

3.1 Representation

Automated techniques to inform social judgments must have some representation of the team exercise including task knowledge (e.g., tasks, their effects, alternative courses of actions), state knowledge (e.g., information from sensors and communication events) and social knowledge (e.g., power relationships between individuals and social obligations). We build on standard representations that have been developed by planning and communication researchers.

Tasks consist of a set of steps, each of which is either a primitive action (e.g., a physical or sensing action in the virtual world) or an abstract action. Abstract actions may be decomposed hierarchically in multiple ways and each decomposition consists of a sequence of abstract or primitive sub-actions. Interdependencies among steps are represented as a set of causal links and threat relations. Each causal link specifies that an effect of a step achieves a particular goal that is a precondition for another step. For example, marking a landing zone with smoke achieves the goal of visually identifying the landing zone for the helicopter, which is a precondition for landing it.

Threat relations specify that an effect of a step threatens a causal link by making the goal unachievable before it is needed. For example, extinguishing the smoke before the helicopter arrives threatens its ability to land.

In addition to understanding the structure of tasks, agents must understand the roles of each team member. Each task step is associated with the team member that performs it, the *performer*. In addition, each task step may be annotated with the teammate who has *authority* over its execution: the performer of a task step cannot execute it until authorization is given by the specified teammate with authority. This requires modeling hierarchical organizational structure of social teams, such as in the military.

Finally, we must model the impact of communication events between entities in the simulation. We follow the Trindi project approach to communication management [Larsson and Traum, 2000]. This approach maintains an explicit information state that is updated by dialogue moves. For example, an assertion by a speaker is a dialogue move that has the effect of establishing a commitment by the speaker that the assertion is true. Orders, on the other hand, can only be issued by a superior to a subordinate in the social structure and have as their effect a social obligation for the subordinate to perform the content of the action.

3.2 From theory to computational approach

Given the representation, we can now revisit the question of representing the core conceptual variables underlying Shaver and Weiner's attribution theories.

Causality: Causal knowledge is encoded via hierarchical task representation. Each action consists of a set of preconditions and effects. The desirability of action effects (i.e. effects having positive/negative significance to an agent) is represented by utility values [Blythe, 1999].

A *non-decision node* is an abstract action that can only be decomposed in one way. A *decision node*, on the other hand, can be decomposed in more than one way. In a decision node, an agent needs to make a decision and select among different choices. If a decision node A can be decomposed in different ways $a_1, a_2, ..., a_n$, we call $a_1, a_2, ..., a_n$ *choices* of action A, and $a_1, a_2, ..., a_n$ are called *alternatives* of each other. Clearly, a primitive action is a non-decision node, while an abstract action can be either a non-decision node or a decision node.

Consequences or outcomes (we use these two terms equally in this paper) of actions are represented as a set of primitive action effects. The *consequence set* of an action A can be defined recursively from leaf nodes (i.e. primitive actions) in plan structure to action A as follows. Con-

sequences of a primitive action are those effects with non-zero utility. For an abstract action, if the abstract action is a non-decision node, then the consequence set of the abstract action is the aggregation of the consequences of its sub-actions. If the abstract action is a decision node, we need to differentiate two kinds of consequences. If consequence p occurs among all the choices of a decision node, we call p a *common consequence* of the decision node; otherwise p is a *non-common consequence* of the decision node. Consequence set of a decision node is defined as the set of its common consequences.

Foreseeability refers to an agent's foreknowledge about actions and consequences. We use *know* and *bring-about* to express foreseeability. If an agent knows an action A brings about a consequence p before the execution of A, then the agent foresees A brings about p.

Intentionality: Intention is generally conceived as a commitment to work toward a certain act or outcome. A key question of credit assignment is distinguishing whether an entity intends an act (*act intent*) versus whether it intends specific consequences of the action (*outcome intent*). Most theories argue that outcome intent rather than act intent is the key factor in determining credit/blame. This is illustrated by the *package deal problem* [Bratman, 1990]. Say military planners wish to bomb a weapons factory but a school is placed within the factory. One might assume the planners intend the act of bombing and intend the outcome of destroying the factory but not intend the outcome of destroying the school.

Following the notations in [Grosz and Kraus, 1996], we use *intend-to* and *intend-that* to distinguish act intention and outcome intention. But since our work is applied to richer social context, we extend the meaning of *intend-to* to include indirect cases. One case is that an agent may intend to act, but may not be the actor himself/herself (e.g. by ordering another agent to act). Another case is that an agent may intend to act but is coerced to do so. *Intend-that*, on the contrary, is used in a more restricted manner. Because of the nature of our problem, we always specify intending an outcome of which action.

Coercion: Similar difference exists in act coercion and outcome coercion. An agent may be coerced to act yet not be necessarily coerced to achieve any specific outcome assuming if there are multiple ways to achieve the task. It is important to differentiate being coerced to act and being coerced to achieve consequence(s) of the action, because it is the latter that actually influence our judgment of behavior, and is used to determine praiseworthy/blameworthy agent. We use *coerced-to* and *coerced-that* to distinguish coerced actions and coerced consequences. In the case of outcome coercion (i.e. *coerced-that* is true), the responsible agent for a conse-

quence is the performer of an action or the entity that has authority over the action, but the action may not be the primitive one that directly leads to the outcome.

3.3 The Attribution Process

Social credit attributions are always from a perceiver's perspective. For the purposes of AAR, this could be a global perspective with access to all simulation events and communications if the goal is to inform an exercise controller. Or it could be from the perspective of a key decision maker if the goal is to allow an autonomous agent to answer questions during the AAR process. Since different perceivers may have different goals, different observations, and different knowledge about the world, it may well be the case that for the same situation, different perceivers form different judgments. For example, an agent may not think himself/herself is blameworthy, but the perceiver thinks the agent is.

Nevertheless, the attribution process is general, and applied uniformly to different perceivers. If an action performed by an agent brings about *positive/negative* consequence, and the agent is not coerced to achieve the consequence, then credit/blame is assigned to the *performer* of the action. Otherwise, assign credit/blame to the *authority*. If the authority is in turn coerced, the process needs to trace further up the hierarchy to find the responsible agent for the consequence.

Coercion is used to determine the praiseworthy/blameworthy agent, while intention and foreseeability are used in assigning the degree of praiseworthiness/blameworthiness. We use a simple categorical model of intensity assignment, though one could readily extend the model to a numeric intensity value by incorporating probabilistic rules of inference. If the responsible agent intends a consequence while acting, the intensity assigned is *high*. If the responsible agent does not foresee the consequence, the intensity is *low*.

4 Inference from Communication and Plans

Judgments of causality, forseeablity, intentionality and coercion are informed by evidence extracted from communication events and from task knowledge. We have developed a number of commonsense rules that allow an automated system to make inferences based on this evidence.

4.1 Inferring from Communication Events

Group simulations often explicitly represent communication events between entities and considerable attribution-relevant information can be extracted from this communication. For example, languages like CCSIL [Salisbury, 1995] or KQML [Finin et al., 1994] represent communication between entities in terms of abstract speech acts [Austin, 1962; Searle, 1969]. When a simulation involves natural language communication between human participants,

underlying speech acts can be acquire using natural language processing technology.

We assume communication between agents is *grounded* [Traum, 1994] and agents communicate *sincerely* and *relevantly* in conversation [Grice's maxims, 1975]. Background information (e.g. agents' social roles, relationship, etc) is also important, for example, an order can be successfully issued only to subordinates; but a request can be made of any agent.

When a speech act is performed, a perceiving agent observes the conversation and makes inferences based on his/her beliefs. As the conversation proceeds, the perceiver acquires new beliefs and updates inferences accordingly.

For our purpose, we are interested in analyzing speech acts that help infer agents' desires, intentions, foreknowledge and choices in acting. We consider the following speech acts (*x* and *y* are different agents. *A* and *B* are actions. *p* is a proposition and *t* is a time stamp):

inform(*x, y, p, t*): *x* informs y that *p* at time *t*.
order(*x, y, A, t*): *x* orders y to do *A* at time *t*.
request(*x, y, A, t*): *x* requests *y* to do *A* at time *t*.
accept(*x, A, t*): *x* accepts to do *A* at time *t*.
reject(*x, A, t*): *x* rejects to do *A* at time *t*.
counterpropose(*x, A, B, y, t*): *x* counters *A* and proposes *B* to *y* at time *t*.

We have identified several commonsense inference rules that allow perceiving agents to form inferences from communication patterns. These rules are general, so can be combined flexibly and applied to variable communication sequences of multiple participants. Here we illustrate two to give a flavor of the approach.

An *order* or a *request* gives evidence that the speaker *desires* the listener to act:

order(x, z, A, t1) $\land$ t1<t2 $\land$ ¬($\exists$t3)(t1<t3<t2 $\land$ believe(x, ¬want(y, do(z, A)), t3)) => believe(x, want(y, do(z, A)), t2)

request(y, z, A, t1) $\land$ t1<t2 $\land$ ¬($\exists$t3)(t1<t3<t2 $\land$ believe(x, ¬want(y, do(z, A)), t3)) => believe(x, want(y, do(z, A)), t2)

The listener may *accept, reject* or *counter-propose* an order/request. Various inferences can be made depending on the response and the power relationship between the speaker and the listener. For instance, if the listener accepts an act wanted by a superior, there is evidence of coercion, and the speaker is viewed as the coercing agent of the action.

believe(x, want(y, do(z, A)), t1) $\land$ accept(z, A, t2) $\land$ t1<t2<t3 $\land$ ¬($\exists$t4)(t2<t4<t3 $\land$ believe(x, ¬intend-to(z, A), t4)) => believe(x, intend-to(z, A), t3)

believe(x, want(y, do(z, A)), t1) $\land$ accept(z, A, t2) $\land$ superior(y, z) $\land$ t1<t2<t3 $\land$ ¬($\exists$t4)(t2<t4<t3 $\land$ believe(x, ¬coerced-to(z, A), t4)) => believe(x, coerced-to(z, A), t3)

Because being coerced to act implies intending to act, the two forms of the rule are consistent. The second form is more specific and thus overrides the first one if both are activated.

4.2 Inferring from Plans

Conversational dialogs and Speech acts provide information about agents' intentions and choices in acting, i.e. *intend-to* and *coerced-to*, but the attribution process actually uses *intend-that* and *coerced-that* for judgment. So we need to solve the problem of inferring outcome intention and outcome coercion from act intention and act coercion.

Different agent may have access to different plans in memory. While plans are specific to certain domain, the structure of plans can be described using domain-independent terms such as action types, alternatives and action effects. To solve the problem in a general way, we make use of the hierarchical task structures, by differentiating action types, comparing consequences of alternatives, and separating common consequences of an action from its non-common ones.

Inferring Outcome Intents from Act Intents

A key question in assigning blame is distinguishing whether an entity intends an act (act intent) and whether the entity intends certain consequences and side effects of that action (outcome intent). A number of rules of evidence can assess these distinctions by considering the task structure. For example, if an entity intends to perform an act, the entity must intend to achieve (at least) one consequence of the action. If the action has only one consequence, then the entity must intend that consequence. In more general cases, when an action has multiple consequences, to infer an entity's intention in achieving a particular outcome, a perceiver may examine alternative acts the agent intends and does not intend, and compare the consequences of intended and unintended alternatives.

We have developed a number of rules of evidence to distinguish *intend-to* from *intend-that*. For example, consider the case illustrated in Figure 4 where there are two alternatives ways to decompose an act (one-squad-forward or two-squads-forward). From the sergeant's perspective, these alternatives have similar consequences with the exception that two squads-forward has an additional consequence that the sergeant considers bad. If in particular, there is only one non-shared consequence *p* of *A* that does not occur in the consequence set of *B*, the agent must intend *p*.

believe(x, intend-to(y, A)) $\land$ believe(x, ¬coerced-to(y, A)) $\land$ believe(x, ¬intend-to(y, B)) $\land$ believe(x, alternative(A, B)) $\land$ believe(x, consequence(B)⊂consequence(A)) => $\exists$p(believe(x, p∈consequence(A)) $\land$ believe(x, p∉consequence(B)) $\land$ believe(x, intend-that(y, p, A)))

As one makes such inference, solutions to inferring outcome intents are partial depending on the information available and comparative features of consequence sets of alternatives.

Inferring Outcome Coercion from Act Coercion
In a non-decision node, if an agent is coerced to act, the agent is also coerced to achieve the consequence of subsequent actions, for the agent has no other choice.

believe(x, coerced-to(y, A)) ∧ believe(x, non-decision-node(A)) ∧ believe(x, p∈ consequence(A)) => believe(x, coerced-that(y, p, A))

In a decision node, however, an agent must make decision amongst multiple choices. Even if an agent is coerced to act, it does not follow that the agent is coerced to achieve certain consequence of the subsequent actions. In order to infer *coerced-that* from *coerced-to* in a decision node, we can examine the choices of the decision node. If an outcome is a common consequence of the node, then it is unavoidable: *coerced-that* is true. Otherwise, if an outcome is a non-common consequence of the node, which means the agent has option to choose the alternative that avoids this outcome, then *coerced-that* is false. Our definition of consequence set ensures the consistency when the rules are applied to the nodes of different levels in plan structure.

Back-Tracing Algorithm
Judgment of attributions is made after the fact (i.e. actions have been executed and the consequence has occurred). The evaluation process starts from the primitive action that directly causes a consequence with positive or negative utility. Since coercion may occur in more than one level in hierarchical plan structure, the process must trace from the primitive action to the higher-level actions. We use a back-tracing algorithm to determine the responsible agent. The algorithm takes as input some desirable or undesirable consequence of a primitive action and works up the task hierarchy. During each pass through the main loop the algorithm uses dialogue and plan evidence rules to assign

---

Algorithm (consequence, plan):
1. *parent=A, where effect of action A is consequence*
2. DO
    2.1 node=parent
    2.2 coerced-to(performer(node), node)=unknown
        coerced-that(performer(node),consequence, node)=unknown
        responsible(consequence)=performer(node)
    2.3 Search dialog history on *node* and apply *dialog rules*
    2.4 IF *coerced-to(performer(node), node)*
    2.5 THEN apply *inference rules* on *node*
    2.6   IF *coerced-that(performer(node), consequence, node)*
    2.7   THEN *responsible(consequence)=authority(node)*
    2.8 parent=P, where P is parent of node in plan
    WHILE *parent≠root* of *plan* AND *coerced-that(performer(node), consequence, node)* is *true*
3. RETURN *responsible(consequence), node*

Figure 3. Back-Tracing Algorithm

---

attributions at the current level (*2.3*). If there is evidence that the performer was coerced to act (*2.4*), rules of evidence also assess outcome coercion (*2.5*). If there is outcome coercion (*2.6*), the authority is deemed responsible (*2.7*). If current action is not the root node in plan structure and outcome coercion is true, the algorithm enters next loop and evaluates the next level up in the task hierarchy.

After the execution of the algorithm, the responsible agent for the outcome is determined. The algorithm may also acquire values of act intention and foreknowledge meanwhile. Then rules for inferring outcome intents can be applied to further determine the responsible agent's intention in achieving the evaluated consequence.

5 Illustration

We illustrate the process of credit assignment through the MRE example introduced above. We focus on three social actors, the *lieutenant*, the *sergeant* and the *squad leader*, who work as a team in task performance. The lieutenant is a human trainee and acts as authority over the sergeant. The squad leader acts as a subordinate of the sergeant. Communications between agents are represented via speech acts and a conversation history as in the MRE.

Take the sergeant's perspective as an example. The sergeant perceives the conversation between the actors and task execution. Conversation history includes the following acts, ordered by the time speakers addressed them. *lt*, *sgt* and *sld* stand for the lieutenant, the sergeant and the squad leader. t1<t2<…<t6.

1. order(lt, sgt, two-squads-fwd, t1)
2. inform(sgt, lt, bring-about(two-squads-fwd, unit-fractured, t2)
3. counter-propose(sgt, lt, two-squads-fwd, one-squad-fwd, t3)
4. order(lt, sgt, two-squads-fwd, t4)
5. accept(sgt, two-squads-fwd, t5)
6. order(sgt, sld, 1st-and-4th-forward, t6)
   … …

Figure 4 illustrates a partial plan the sergeant has access to. *One squad forward* and *two squads forward* are two choices of action *support Eagle 1-6. One squad forward* is composed of two primitive actions, *4th squad recon (forward)* and *remaining (squads) forward. Two squads forward* consists of *1st and 4th (squads) forward* and *2nd and 3rd (squads) forward*. Effects of primitive actions are shown under the graph. Assume two effects are salient to the sergeant. *1-6 supported* is a desirable team goal. The sergeant assigns negative utility to *unit fractured* and that this consequence serves as input to the back-tracing algorithm. We illustrate how to find the blameworthy agent given the sergeant's knowledge and observations of communication events and task execution.

Loop 1: The algorithm starts from primitive action *1st-and-4th-fwd*, of which *unit fractured* is an effect. The sergeant perceived the action was executed by the squad leader.
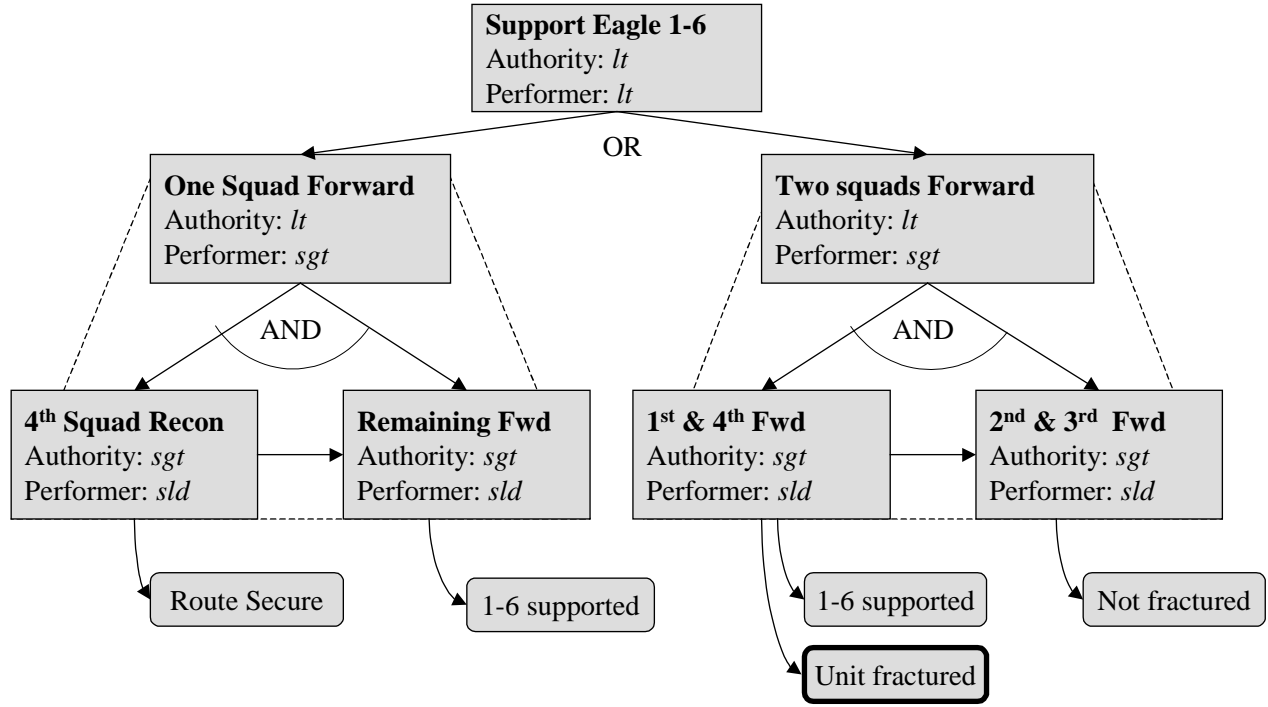
Figure 4: Team Plan from Sergeant's Perspective

Step 2.2: By default, *coercd-to*(*sld, 1st-and-4th-fwd*) is unknown, *coerced-that*(*sld, unit-fractured, 1st-and-4th-fwd*) is unknown. Assign the squad leader to the responsible agent.

Step 2.3: Relevant conversation history is act 6. Since the sergeant ordered the squad leader the action, through inference rules the algorithm infers that the sergeant believes that he wants the squad leader to act. Since the squad leader accepted by executing the action and the sergeant is the superior, the algorithm infers that the sergeant believes that he coerced the squad leader to act.

Step 2.4–2.5: Since *coerced-to* is true and the primitive action is a *non-decision node*, then through the application of an inference rule, the sergeant believes he coerced the squad leader to fracture the unit.

Step 2.6–2.7: Since *coerced-that* is true, assign the coercing agent to the responsible agent. The sergeant believes he is responsible for *unit-fractured*.

Since parent node is not the root in plan structure and *coerced-that* is true, the algorithm enters next loop. We leave the executions of subsequent loops to the reader. The main results are given below.

Loop 2: The action is *two-squads-fwd*, performed by the sergeant. Relevant conversation history is sequence 1−5. A variety of beliefs can be inferred from commonsense rules by analyzing the task structure and communication history.

believe(sgt, want(lt, do(sgt, two-squads-fwd)))
believe(sgt, know(lt, bring-about(two-squads-fwd, unit-fractured)))
believe(sgt, know(lt, alternative(one-squad-fwd, two-squads-fwd)))
believe(sgt, intend-to(lt, two-squads-fwd))
believe(sgt, ¬intend-to(lt, one-squad-fwd))
believe(sgt, coerced-to(sgt, two-squads-fwd))
believe(sgt, coerced-that(sgt, unit-fractured, two-squads-fwd))

After *loop 2*, the sergeant believes the lieutenant coerced him to fracture the unit. So the lieutenant is responsible for the outcome.

Loop 3: The action is *support-Eagle 1-6*, performed by the lieutenant. There is no relevant conversation in history. The values of *coerced-to*, *coerced-that* and responsible agent are default. There is no clear evidence of coercion, so the sergeant believes that the lieutenant is the responsible agent. Parent node is the root in plan structure. The algorithm terminates.

Now the sergeant also has the belief that the student intended to send two squads forward and did not intend to send one squad forward (acquired in *loop 2*). Since the consequence set of *one-squad-fwd* (i.e. *1-6-supported*) is

subset of the consequence set of *two-squads-fwd* (i.e. *1-6-supported* and *unit-fractured*), through inference rules the algorithm determines the sergeant believes that the lieutenant intended *unit-fractured*. So the lieutenant is to *blame* with *high* intensity.

6   Conclusions

A key element of the after action review process is to form social explanations of blame or credit and to allow various individuals to compare and contrast their individual explanations in order to arrive at a more objective understanding of the group exercise. In this article we present a preliminary computational approach to automate this social credit assignment process. Two obvious applications of this model for AAR are in allowing synthetic agents to participate in the AAR process (by forming and relating their own social attributions) and by assisting a human AAR facilitator by automatically processing a global simulation trace and identifying key events or decisions that need discussion.

The problem of social credit assignment is central in social psychology and social cognition. With the development of human-like agent systems, it is increasingly important for computer-based systems to model this human-centric form of social inference. Our work attempts to help bridge between psychological accounts and computational models by means of AI methods. Rather than impose arbitrary rules on judgment process, our work relies on commonsense heuristics of human inference from communication events and plans as knowledge states of agents. Our treatments are domain-independent, thus can be used as a general approach to the problem.

In the future, we must incorporate probabilistic reasoning to deal with uncertainty in observations and judgment process. For modeling more complex multi-agent teamwork, we need to consider joint responsibility and sharing responsibility among teammates (the current model assumes one agent has sole responsibility). The inferences related to foreseeability are too restrictive and need to make better use of plan knowledge, specifically considering how actions may be coerced not just directly through orders, but indirectly due to the effects of actions of other entities. For example, an action taken by one entity may, with foreknowledge, coerce another entity to pursue a less desirable option. As our task representations already encode information about action preconditions and effects, this type of inference is a natural extension of our existing methods.

Acknowledgement

References

J. Austin. *How to Do Things with Words*. Harvard University Press, 1962.

J. Blythe. Decision-Theoretic Planning. *AI Magazine*, 20(2):37-54, 1999.

M. Bratman. *Intention, Plans and Practical Reason*. Harvard University Press, Cambridge, Massachusetts, 1987.

T. Finin, R. Fritzson, D. McKay and R. McEntire. KQML as an Agent Communication Language. *Proceedings of the 3rd International Conference on Information and Knowledge Management,* ACM Press, 1994.

J. Gratch, J. Rickel, E. André, N. Badler, J. Cassell and E. Petajan. Creating Interactive Virtual Humans: Some Assembly Required. *IEEE Intelligent Systems*, 17(4), July/August, pp. 54-63, 2002.

H. P. Grice. Logic and Conversation. In: P. Cole and J. Morgan (Eds.), *Syntax and Semantics: Vol 3 Speech Acts*. Academic Press, 1975.

B. Grosz and S. Kraus. Collaborative Plans for Complex Group Action. *Artificial Intelligence*, 86(2): 269-357, 1996.

W. L. Johnson. Agents that Learn to Explain Themselves. *Proceedings of the National Conference on Artificial Intelligence*, Seattle, WA, August, 1994.

S. Larsson and D. Traum. Information state and dialogue management in the TRINDI dialogue move engine toolkit. *Natural Language Engineering,* 6:323-340, September 2000, Special Issue on Spoken Language Dialogue System Engineering.

L. McCarty. Some Arguments about Legal Arguments. In: *Proceedings of International Conference on Artificial Intelligence and Law*, Melbourne, 1997.

M. E. Pollack. Plans as Complex Mental Attitudes. In: P. Cohen, J. Morgan and M. E. Pollack (Eds.), *Intentions in Communication*. MIT Press, 1990.

J. Rickel, S. Marsella, J. Gratch, R. Hill, D. Traum and B. Swartout. Toward a New Generation of Virtual Humans for Interactive Experiences. *IEEE Intelligent Systems*, 17(4), July/August, pp.32-38, 2002.

A. Newell. *Unified Theories of Cognition.* Harvard University Press, 1990.

M. R. Salisbury. Command and Control Simulation Interface Language (CCSIL): Status Update. *Twelfth Workshop on Standards for the Interoperability of Defense Simulations*. Orlando Florida. 1995.

J. R. Searle. *Speech Acts*. Cambridge University Press, Cambridge, 1969.

K. G. Shaver. *The Attribution Theory of Blame: Causality, Responsibility and Blameworthiness*. Springer-Verlag, 1985.

D. Traum. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. Thesis, University of Rochester, 1994.

B. Weiner. *The Judgment of Responsibility*. Guilford Press, 1995.

Author Biographies

JONATHAN GRATCH heads up the stress and emotion project at the University of Southern California's Institute for Creative Technology and a research assistant professor with the computer science department at USC. He has worked for a number of years in the areas of simulation, cognitive science, planning and machine learning. He received his Ph.D. from the University of Illinois in 1995.

WENJI MAO is a Ph.D. student at the University of Southern California and research assistant at USC's Institute for Creative Technologies. She has worked in the areas of emotion modeling and intelligent agents. She received her M.S. from Chinese Academy of Sciences.